Lee, Yoonkyung; Wang, Rui

Does modeling lead to more accurate classification? A study of relative efficiency in linear classification.


Summary: Classification arises in a wide range of applications. A variety of statistical tools have been developed for learning classification rules from data. Understanding of their relative merits and comparisons help users to choose a proper method in practice. This paper focuses on theoretical comparison of model-based classification methods in statistics with algorithmic methods in machine learning in terms of the error rate. Extending Efron's comparison of logistic regression with linear discriminant analysis (LDA) under the normal setting, we contrast such algorithmic methods as the support vector machine (SVM) and boosting with the LDA and logistic regression and study their relative efficiencies in reducing the error rate based on the limiting behavior of the classification boundary of each method. We show that algorithmic methods are generally less effective than model-based methods in the normal setting. In particular, loss of efficiency in error rate is typically about 33% to 60% for the SVM and 50% to 80% for boosting when compared to the LDA. However, a smooth variant of the SVM is shown to be even more efficient than logistic regression. In addition to the theoretical study, we present results from numerical experiments under various settings for comparisons of finite-sample performance and robustness to mislabeling and model misspecification.

Keywords: boosting; classification; efficiency; error rate; LDA; logistic regression; mislabeling; robustness; SVM
doi:10.1016/j.jmva.2014.09.010